Protein Modeling

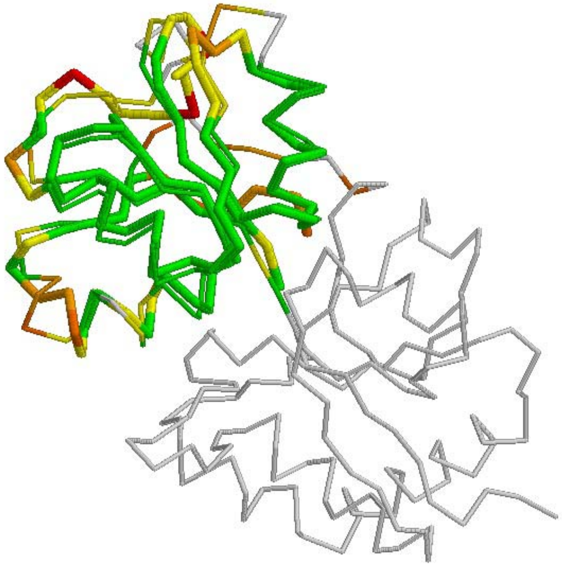
Adam Zemla, Carol Zhou

Protein Modeling in Support of Biodefense
March 17, 2004

UCRL-PRES-202984

We will examine the role that structure modeling plays in development of protein signatures…and more.

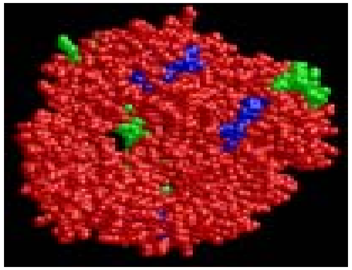
- What are protein signatures & why do we need them?
- How does protein modeling assist us in choosing protein signature targets?
- Why are empirically determined structures not sufficient?
- How will our research advance the field of protein modeling?
- How can our research apply to advances in biology?



What is a protein signature target?

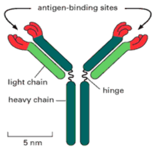
A region for identification of a target protein, which is:

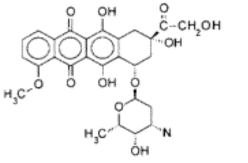
- a specific sequence/fold than can be recognized by a ligand (binder)
- unique to the protein of interest



We identify multiple regions for each pathogen.

These regions can be exploited by a variety of detection chemistries and platforms.





DOXORUBICIN (MW=579.98)

lys-leu-val-thr-pro

Protein signatures allow us to detect:

- pathogens
- proteins associated with virulence or toxicity

There are several reasons why protein signatures are necessary

- Many virus genomes are too variable for other detection methods
    - Adequate conservation exists in protein-space
- Other types of signatures could be “engineered around” to thwart detection
    - Harder to alter proteins without changing function
- Orthogonal confirmation is desired (complement other methods)
- Protein assays could confirm viability

Our protein pipeline leverages structure modeling capabilities

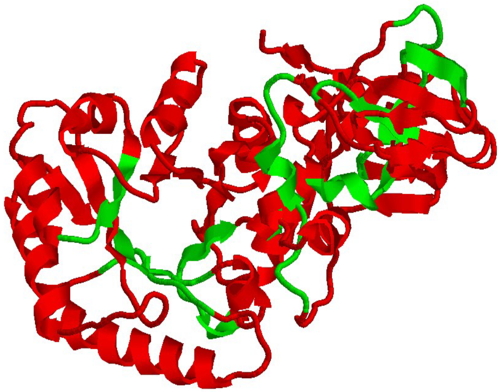
Raw protein sequence

```
MKREIEWNAIIELGVRPMSLKYGRDTIVEVDLNAVKHNVKEFKKRVNDENIAMMAAVKAN
GYGHGAVEVAKAAIEAGINQLAIAFVDEAIELREAGINVPILILGYTSVAAAEEAIQYDV
MMTVYRSEDLQGINEIANRLXKKAQIQVKIDTGMSRIGLQEEEVKPFLEELKRMEYVEVV
GMFTHYSTADEIDKSYTNMQTSLFEKAVNTAKELGIHIPYIHSSNSAGSMEPSNTFQNMV
RVGIGIYGMYPSKEVNHSVVSLQPALSLKSKVAHIKHAKKNRGVSYGNTYVTTGEEWIAT
VPIGYADGYNRQLSNKGHALINGVRVPVIGRVCMDQLMLDVSKAMPVQVGDEVVFYGKQG
EENIAVEEIADMLGTINYEVTCMLDRRIPRVYKENNETTAVVNILRKN
```

Conserved & unique protein sequence

```
MKREIEWNAIIELGVRPMSLKYGRDTIVEVDLNAVKHNVKEFKKRVNDENIAMMAAVKAN
GYGHGAVEVAKAAIEAGINQLAIAFVDEAIELREAGINVPILILGYTSVAAAEEAIQYDV
MMTVYRSEDLQGINEIANRLXKKAQIQVKIDTGMSRIGLQEEEVKPFLEELKRMEYVEVV
GMFTHYSTADEIDKSYTNMQTSLFEKAVNTAKELGIHIPYIHSSNSAGSMEPSNTFQNMV
RVGIGIYGMYPSKEVNHSVVSLQPALSLKSKVAHIKHAKKNRGVSYGNTYVTTGEEWIAT
VPIGYADGYNRQLSNKGHALINGVRVPVIGRVCMDQLMLDVSKAMPVQVGDEVVFYGKQG
EENIAVEEIADMLGTINYEVTCMLDRRIPRVYKENNETTAVVNILRKN
```

3D model showing location of candidate protein signature target



Structural homology provides high-resolution modeling

Annotation selection

Targets have potential use for detection, therapeutics, or vaccines

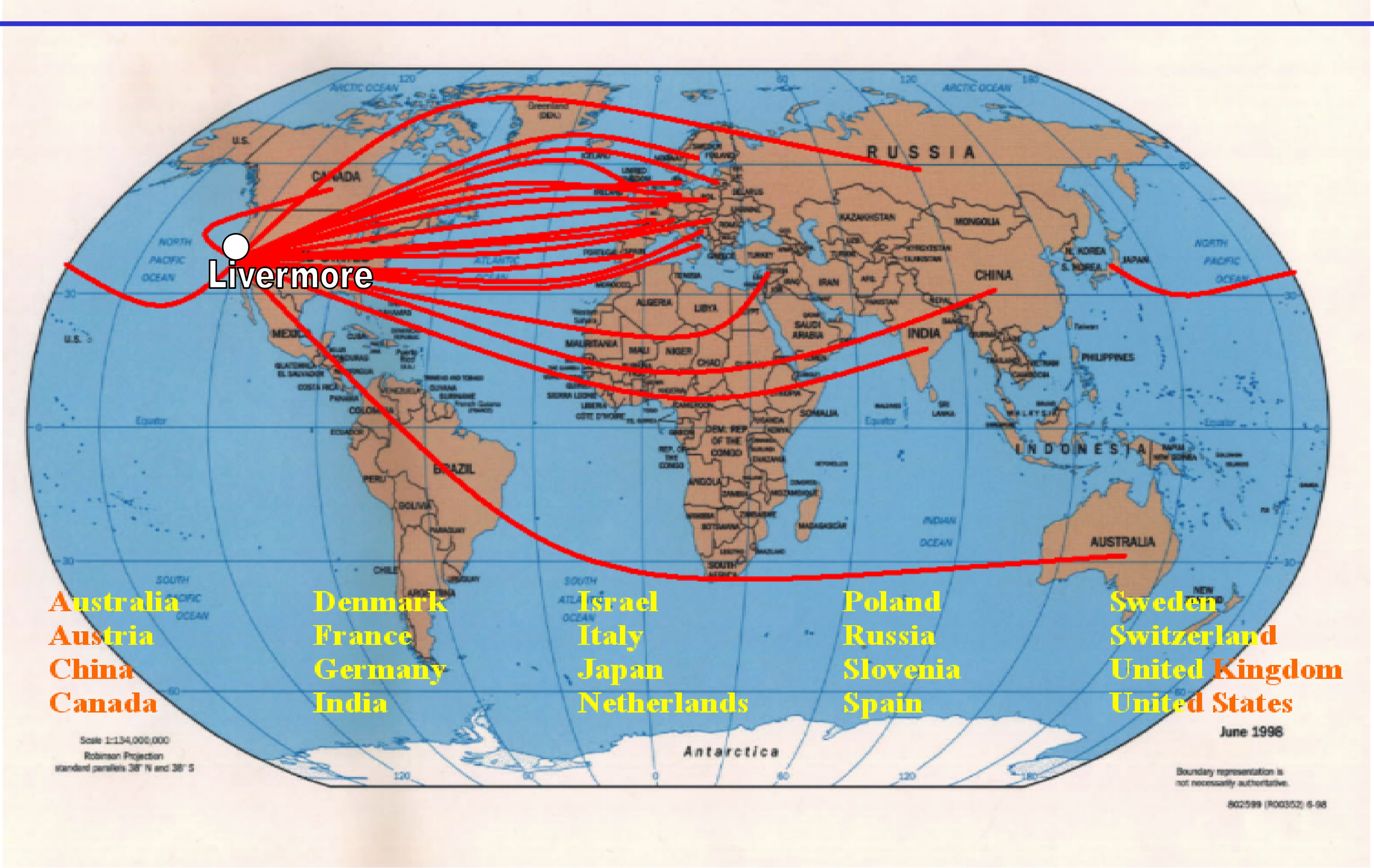
Why is modeling necessary?

- **Number of proteins whose structure and biochemical function are unknown increases exponentially.** Number of proteins (genes) discovered daily: **~1000**
- **Cost and time required to experimentally characterize these new proteins is prohibitive.** Number of daily experimentally determined structures: **~10**
- **Not all proteins can be solved experimentally.**
- **Number (March 09, 2004) of structures deposited in Protein Data Bank (PDB) as of 9 March 2004: 24,615**
- **Current number of folds classified by Structural Classification of Proteins database (SCOP): 800 (out of ~10,000 est'd total)**
- **Computational methods hold great promise in uncovering the structure and function of many new proteins**

Participation in CASP extends worldwide

187 prediction groups in CASP5

28,728 processed models



Australia
Austria
China
Canada

Denmark
France
Germany
India

Israel
Italy
Japan
Netherlands

Poland
Russia
Slovenia
Spain

Sweden
Switzerland
United Kingdom
United States

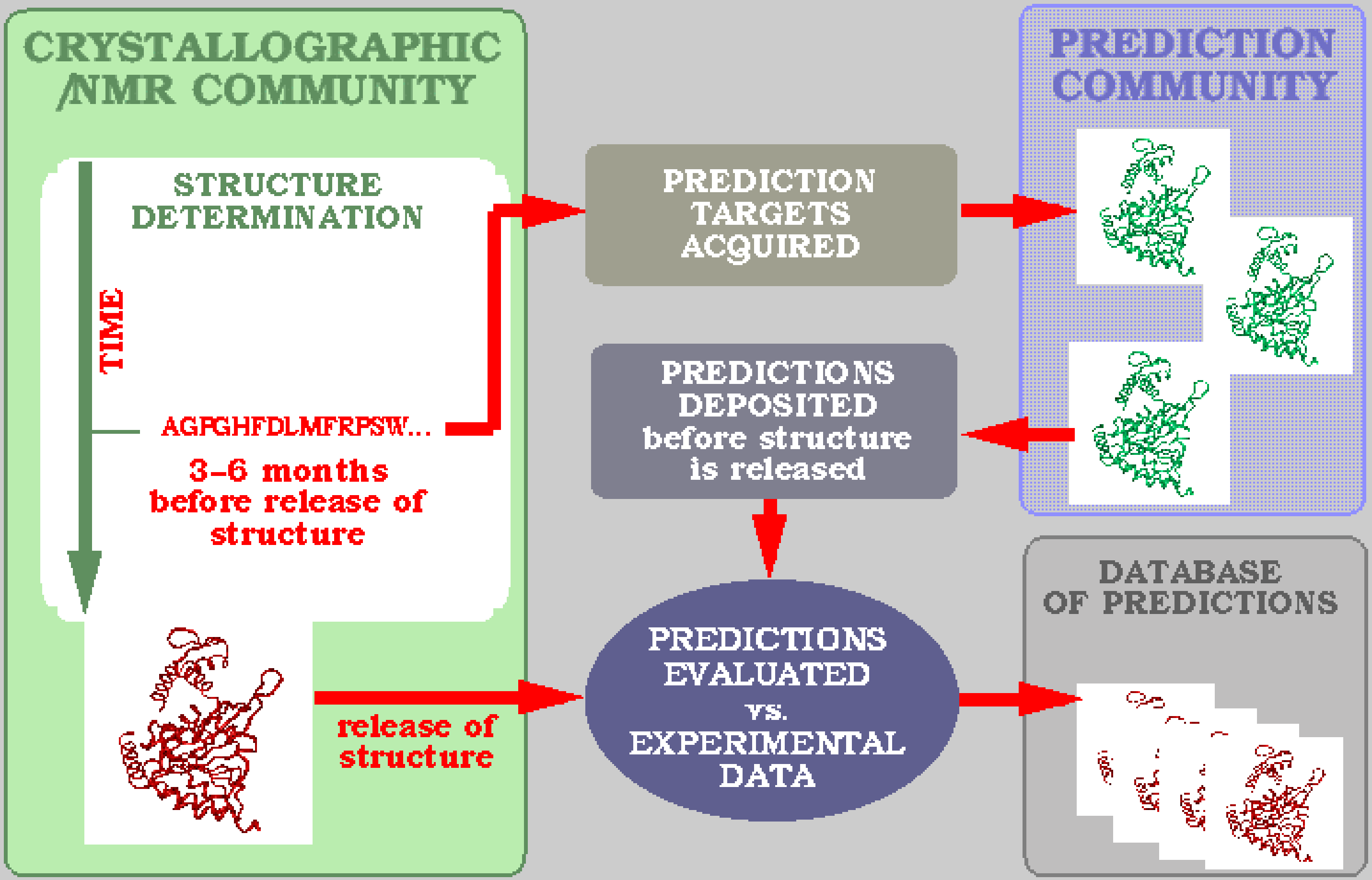
June 1998

Scale 1:134,000,000
Robinson Projection
standard parallels 38° N and 38° S

Boundary representation is
not necessarily authoritative.

802599 (R00352) 6-98

CASP: The blind prediction regime



CRYSTALLOGRAPHIC /NMR COMMUNITY

STRUCTURE DETERMINATION

TIME

AGPGHFDLMFRPSW...

3–6 months before release of structure

release of structure

PREDICTION TARGETS ACQUIRED

PREDICTION COMMUNITY

PREDICTIONS DEPOSITED before structure is released

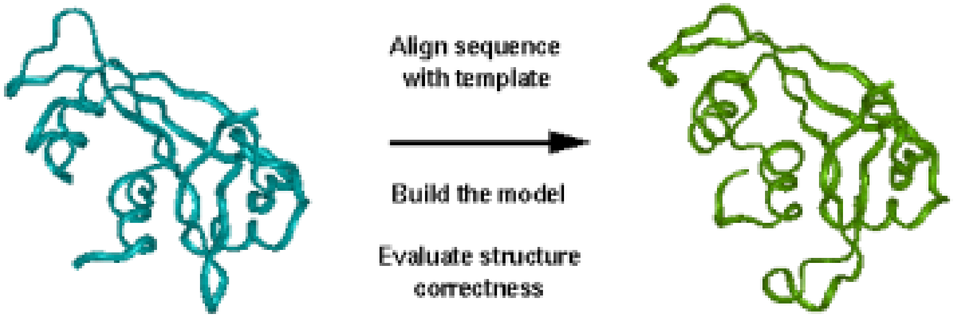
PREDICTIONS EVALUATED vs. EXPERIMENTAL DATA

DATABASE OF PREDICTIONS

CASP: 3 categories of structure prediction

Comparative modeling:

VASFGGQKLTLKKSVITSARRQNDEERIHSTCCLVRDDEQQRAGGGACLVV
VATFAGQKLTLRKTVMTSARKQNEEERIHSTACLVRDDESTMMRGGACIVA

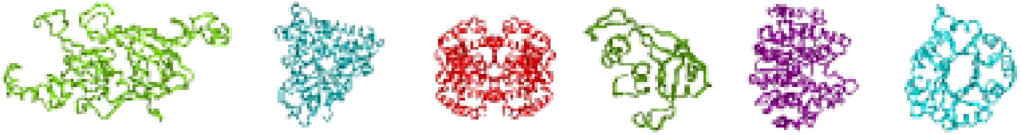


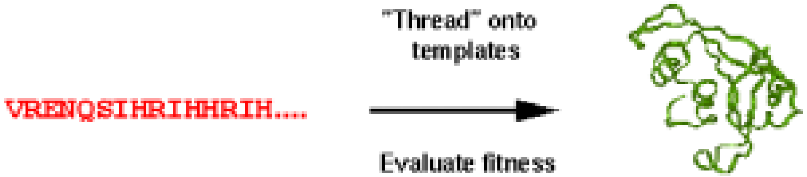
Align sequence with template

Build the model

Evaluate structure correctness

Fold recognition:



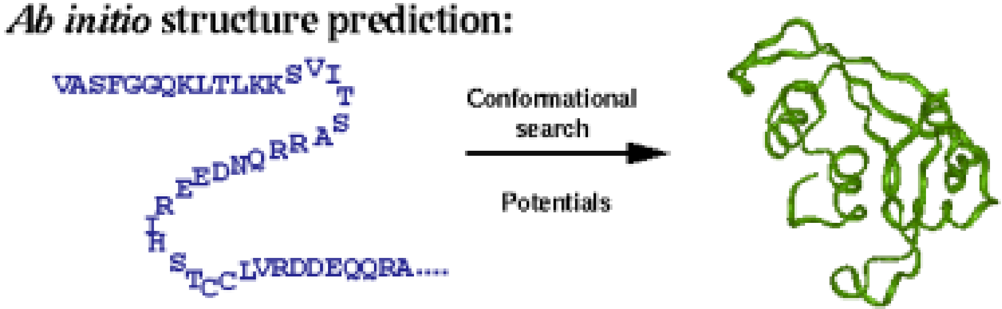


"Thread" onto templates

VRENQSIHRIHHRIH....

Evaluate fitness

Ab initio structure prediction:



VASFGGQKLTLKKSVITSARRQNDEERIHSTCCLVRDDEQQRA....

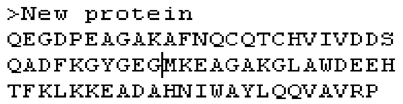
Conformational search

Potentials

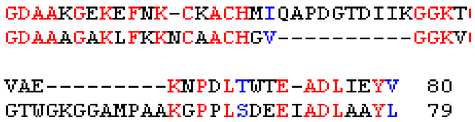
We built an automatic 3D modeler

Main steps in homology modeling:

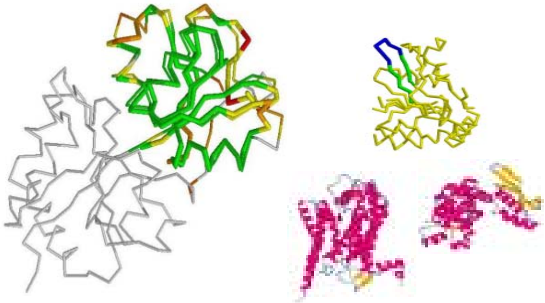
1. Search for similar proteins in Protein Data Bank (PDB) – sequence alignment
2. Verifiy alignments (LGA structure comparison)
3. Build in missing regions (LGA) – “backbone” now complete
4. Add amino-acids (side chains)

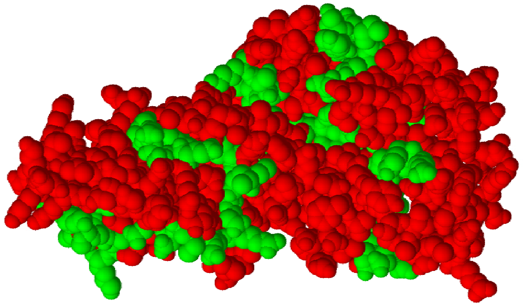


```
>New protein
QEGDPEAGAKAFNQCQTCHVIVDDS
QADFKGYGEGMKEAGAKGLAWDEEH
TFKLKKEADAHNIWAYLQQVAVRP
```

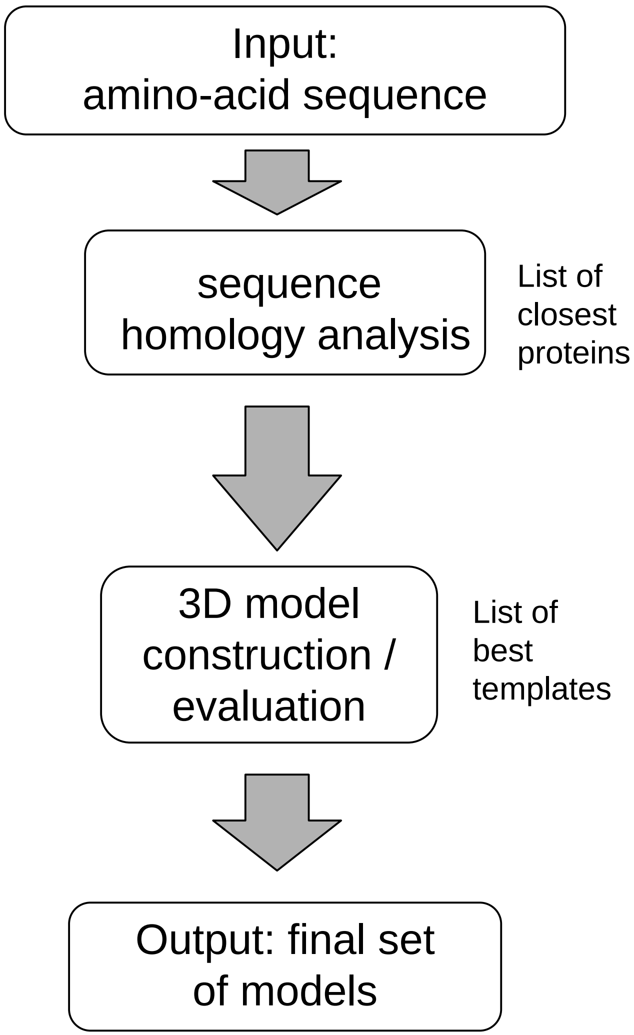


```
GDAAKGEKEFNK-CKACHMIQAPDGTDIIKGGKT
GDAAAGAKLFKKNCAACHGV----------GGKV

VAE----------KNPDLTWTE-ADLIEYV   80
GTWGKGGAMPAAKGPPLSDEEIADLAAYL   79
```





AS2TS server



Input:
amino-acid sequence

sequence
homology analysis

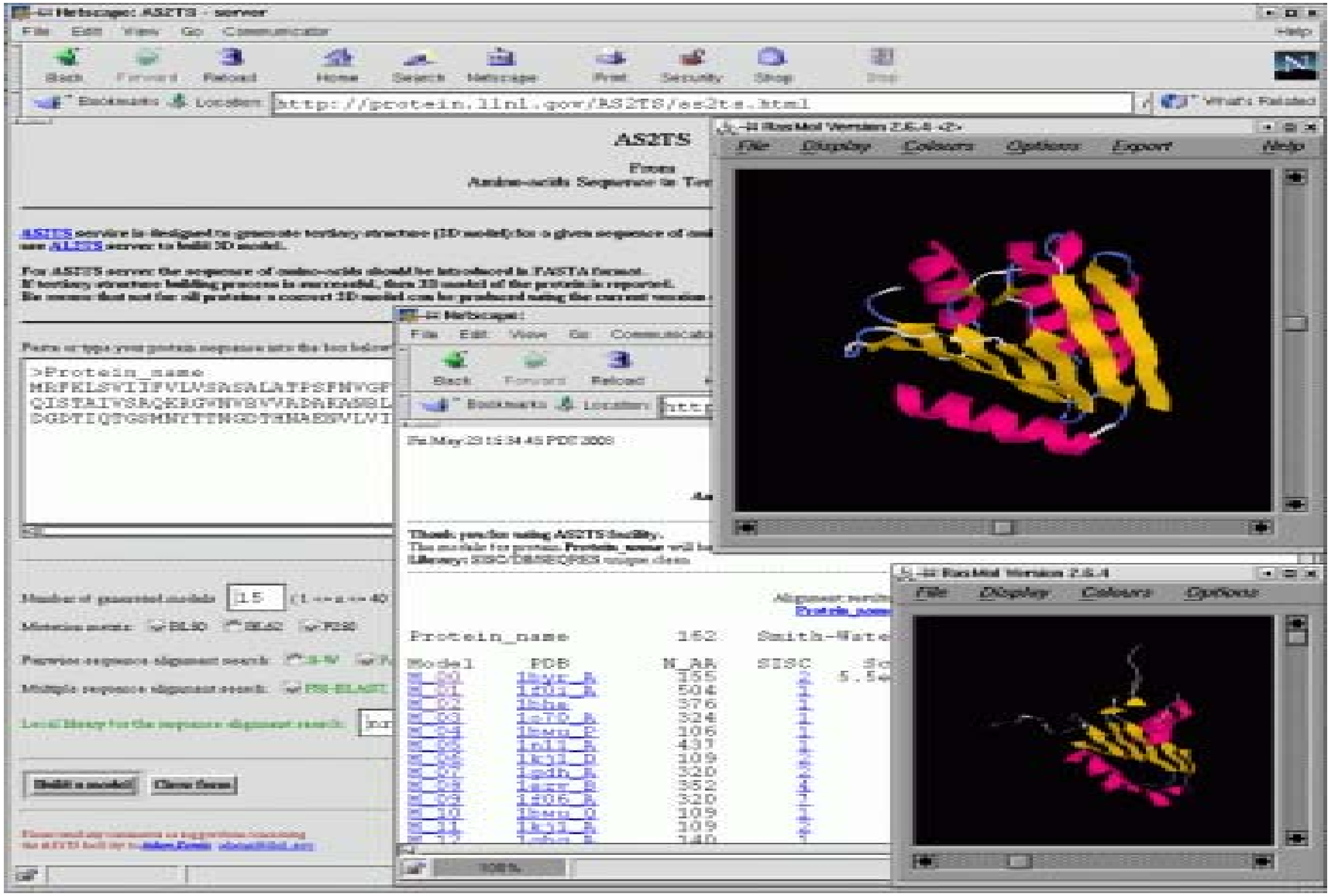
List of
closest
proteins

3D model
construction /
evaluation

List of
best
templates

Output: final set
of models

…and scaled our modeler for whole-proteome analysis



Netscape: AS2TS - server
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: What's Related
RasMol Version 2.6.4 <2>
File Display Colours Options Export Help
Netscape:
File Edit View Go Communicator
Back Forward Reload
Bookmarks Location:
RasMol Version 2.6.4
File Display Colours Options
Build a model
Clear form
100%

http://protein.llnl.gov/AS2TS/as2ts.html

AS2TS

From
Amino-acids Sequence to Te[illegible]

AS2TS service is designed to generate tertiary structure (3D model) for a given sequence of ami[illegible] use AL2TS server to build 3D model.

For AS2TS server the sequence of amino-acids should be introduced in FASTA format.
If tertiary structure building process is successful, then 3D model of the protein is reported.
Be aware that not for all proteins a correct 3D model can be produced using the current version [illegible]

Paste or type your protein sequence into the box below:

```
>Protein_name
MRFKLSVIIFVLVSASALATPSFNVGF
QISTAIVSRQKRGVHVSVVADAKASSL
DGDTIQTGSMNYTTNGDTHNAESVLVI
```

Number of generated models 15 (1 <= n <= 40)

Mutation matrix: BL50 BL62 PZ80

Pairwise sequence alignment search: S-W [illegible]

Multiple sequence alignment search: PSI-BLAST

Local library for the sequence alignment search: [illegible]

[illegible] Adam Zemla [illegible]

Fri May 23 15:34:45 PDT 2003

Thank you for using AS2TS facility.
The models for protein Protein_name will be [illegible]
Library: SISC/DBSEQRES unique chain

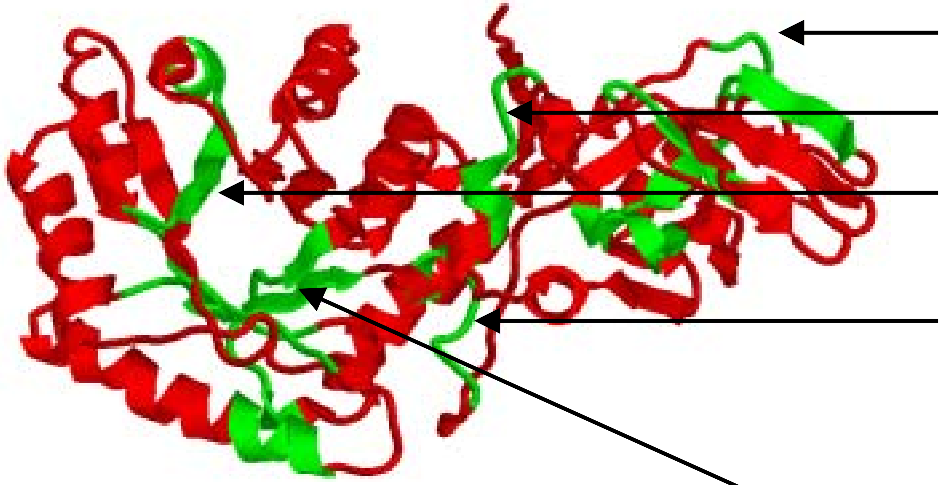
Alignment results
Protein_name

```
Protein_name             152    Smith-Wate
Model        PDB           N_AA     SISC     Sc
M_00         1byr_A        355        2      5.5
M_01         1f01_A        504        1
M_02         1bhe          376        1
M_03         1o7D_A        324        1
M_04         1bwu_P        106        1
M_05         1n11_A        437        1
M_06         1k31_D        109        1
M_07         1gdh_A        320        2
M_08         1azr_B        352        4
M_09         1206_A        320        7
M_10         1bwu_Q        109        1
M_11         1k31_A        109        2
M_12         1[illegible]  140        1
```

A small virus proteome has ~12 proteins, a typical bacterium has 2000

Candidate signature targets can be visualized and selected based on surface accessibility

```
MATPQISRKALASLLLLVAAAAAVSTASADDVLALTESTFEKEVGQDRAALVEFYAPWCGHCKKLAPEYE
KLGASFKKAKSVLIAKVDCDEHKSVCSKYGVSGYPTIQWFPKGSLEPKKYEGQRTAEALAEYVNSEAATN
VKIAAVPSSVVVLTPETFDSVVLDETKDVLVEFYAPWCGHCKHLAPIYEKLASVYKQDEGVVIANLDADK
HTALAEKYGVSGFPTLKFFPKGNKAGEDYDGGRELDDFVKFINEKCGTSRDSKGQLTSEAGIVESLAPLV
KEFLGAANDKRKEALSKMEEDVAKLTGPAAKYGKIYVNSAKKIMEKGSEYTKKESERLQRMLEKGLT
```

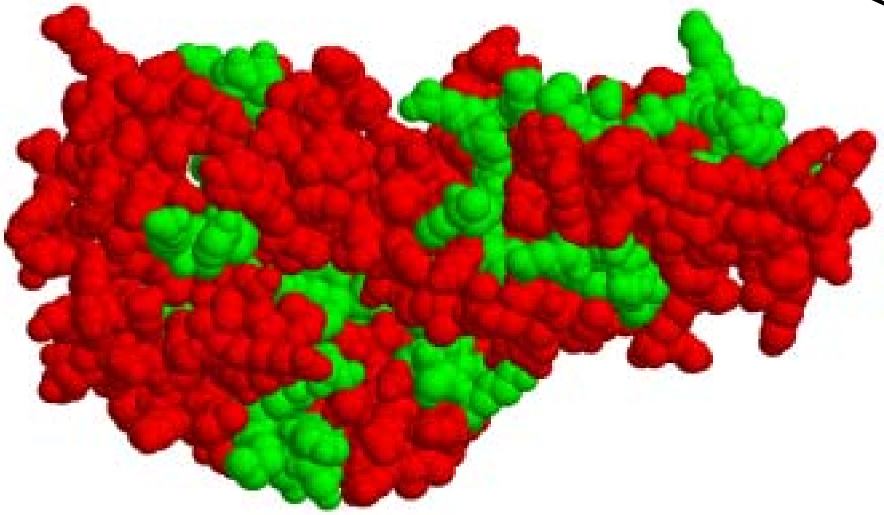


ILVFM

CAGPTSYWQN

HEDKRIVMCGPSW

NEKIVMCGTY



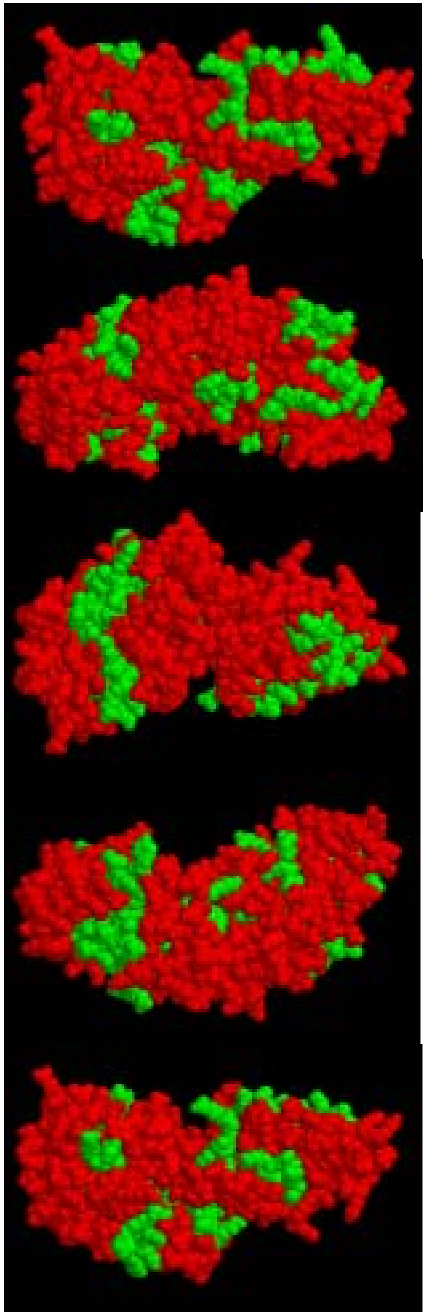
Set of four unique regions inside the beta-sheet barrel:

LMGSQE

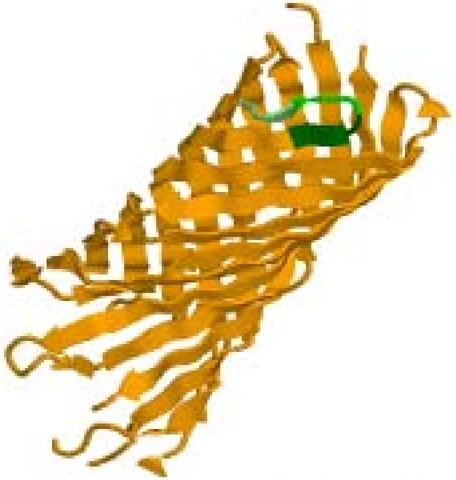
RIFATWHKK

RKDEHN

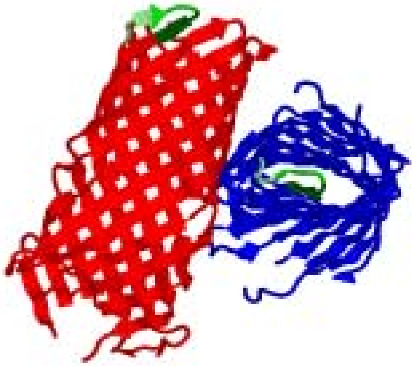
QWYSTPG

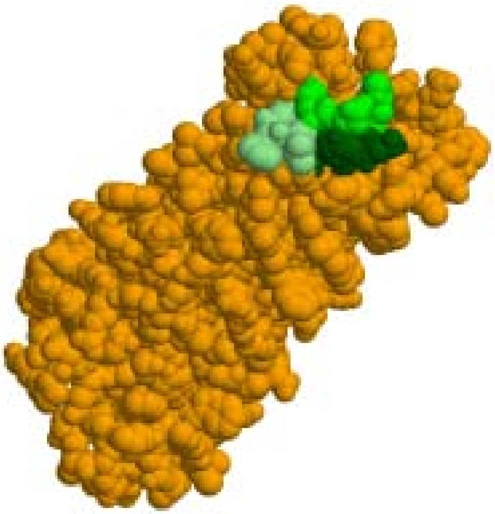


Modeling a protein complex provides additional information



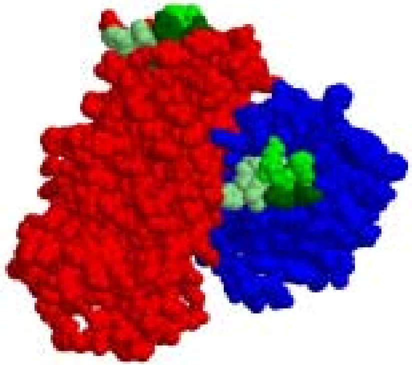
```
MATPQISRKALASLLLLVAAAAAVSTASADDVLALTESTFEKEVGQ
KLGASFKKAKSVLIAKVDCDEHKSVCSKYGVSGYPTIQWFPKGSLE
VKIAAVPSSVVVLTPETFDSVVLFMCEDKCGTWCGHCKHLAPIYEK
HTALAEKYGVSGFPTLKFFPKGNKAGEDYDGGRELDDFVKFINEKC
KEFLGAANDKRKEALSKMEEDVAKLTGPAAKYGKIYVNSAKKIMEK
```





Two overlapping unique regions
131-EDKCGT-136
128-FMCEDK-133

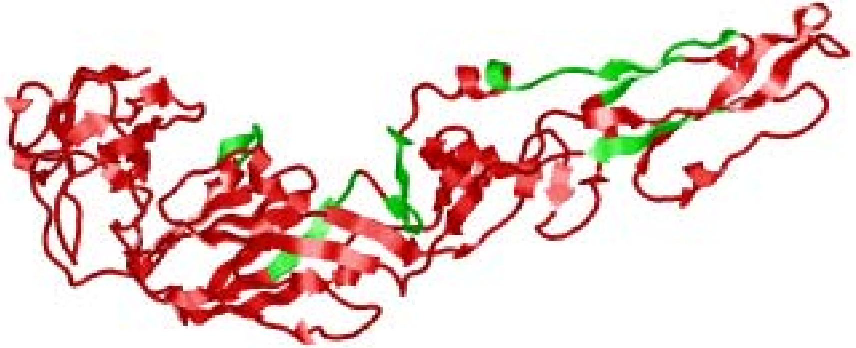
located on the loop on the top of the vase-shaped beta-barrel

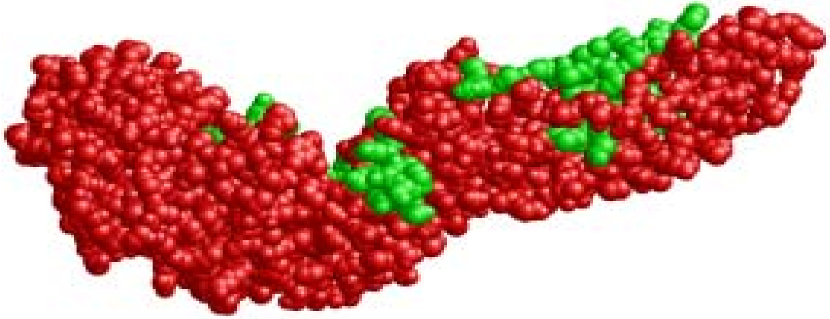


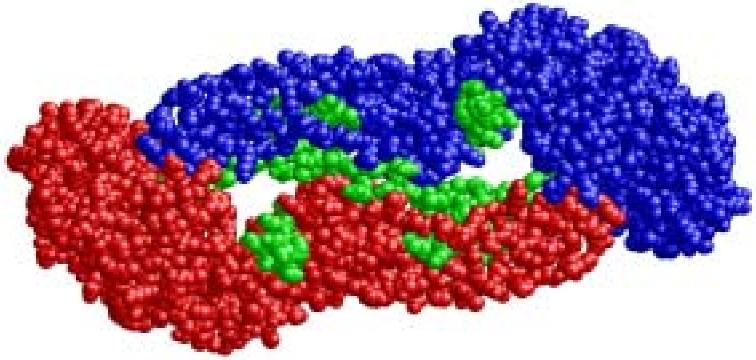
Some signature targets are shielded in the complex

```
-N-LGMSNR-F--G--GA-NV-L--E--S-V--H-KDK-T-DVKHM-HE--N-A--RSYC----V--LST
-A-C--M-E--ND--A-F-FV--QG-V---NG-----F-KGS-DTCAKFA--TKA-G-----E--K--V-
-F-HGP--V-S---Y-TQ----Q-GR--I-----SY----G-YG---VDCEP-SG-DT--YYVMT--TK-
-LV-REN-MDLNL-NSS-G---N-NRET-MEFEEPH--K-SVIALGSQKG-LHQA--GA-PVEFSSNTVK
--SGHLKC---------LK-T--GVCSK-FK-L-TPAD----T-VLE-QY-GTDG-CKV-ISS-ASL-DL-
PVG-L--VN-F----TA---VLI--EPFFG--YIV--RG-QQ----W-KSG-SIGK--T--L---QR---
-G-TAMDFG-VG-V-T---KA-HQ---GA-RS--G-H-NI----LG----WM----R--SI-LT---VG-
--L-LSVNV--
```

3D model based on homology to the envelope glycoprotein from TICK-BORNE ENCEPHALITIS virus (1svb from PDB) described as a flat, elongated dimer, being a component of the complete E protein which would lie on the surface of the viral membrane.







3D model of dimer (chain A in red, chain B in blue, signature regions in green)

West Nile Virus glycoprotein [strain RO97-50]

CONSERVED and UNIQUE signature regions (at least 6 residues long)

Structure modeling remains an imperfect science

Homology modeling produces:

- good models for 30-40% of proteins
- fair models for another ~30%

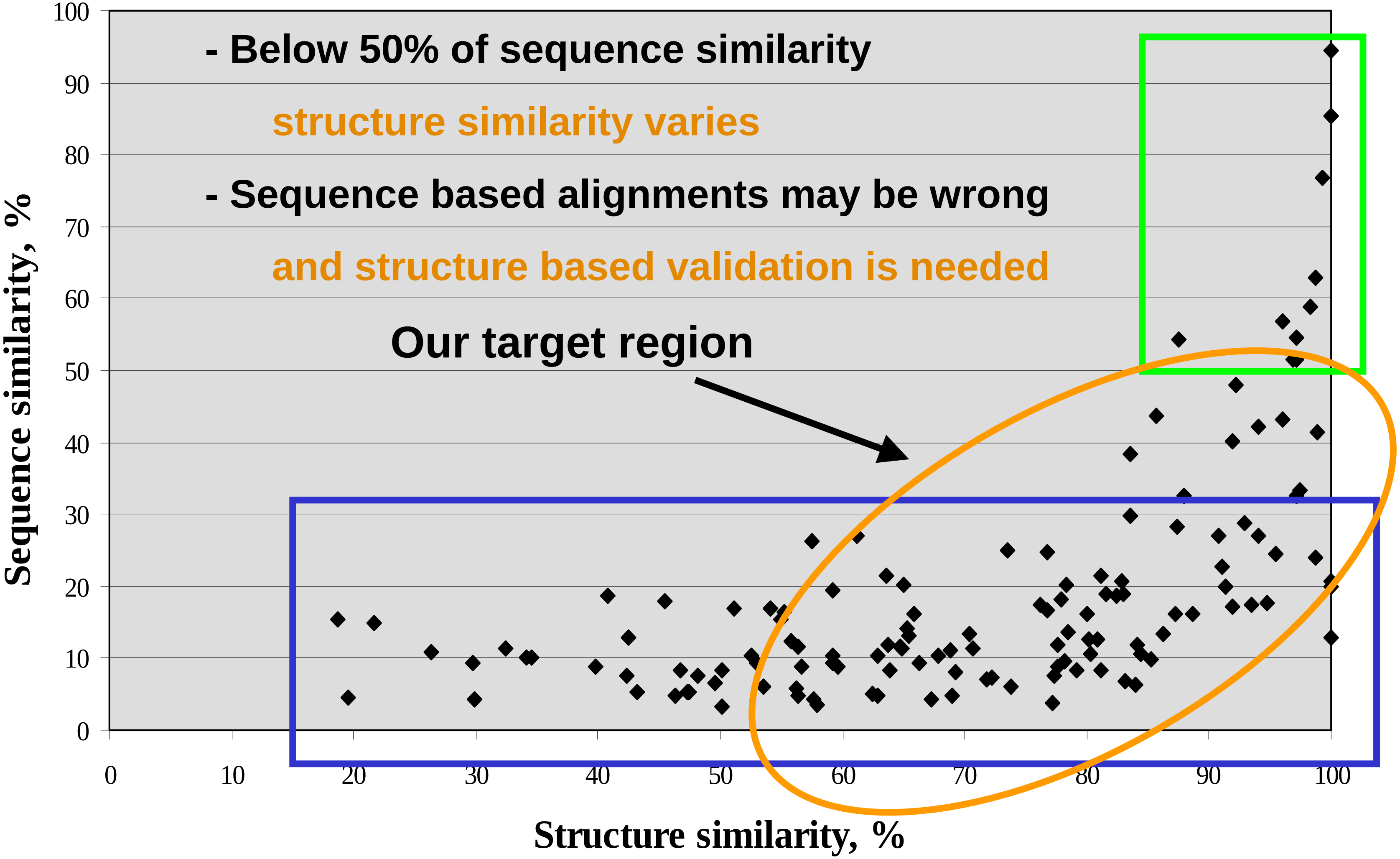
Homology modeling is useful for

- high-throughput, whole-proteome screening
- candidate signature target selection

More work is needed to:

- develop methods for protein structure comparison
- define new structural folds
- classify proteins based on structure correspondence

Structure similarity is more conserved than sequence similarity



- Below 50% of sequence similarity

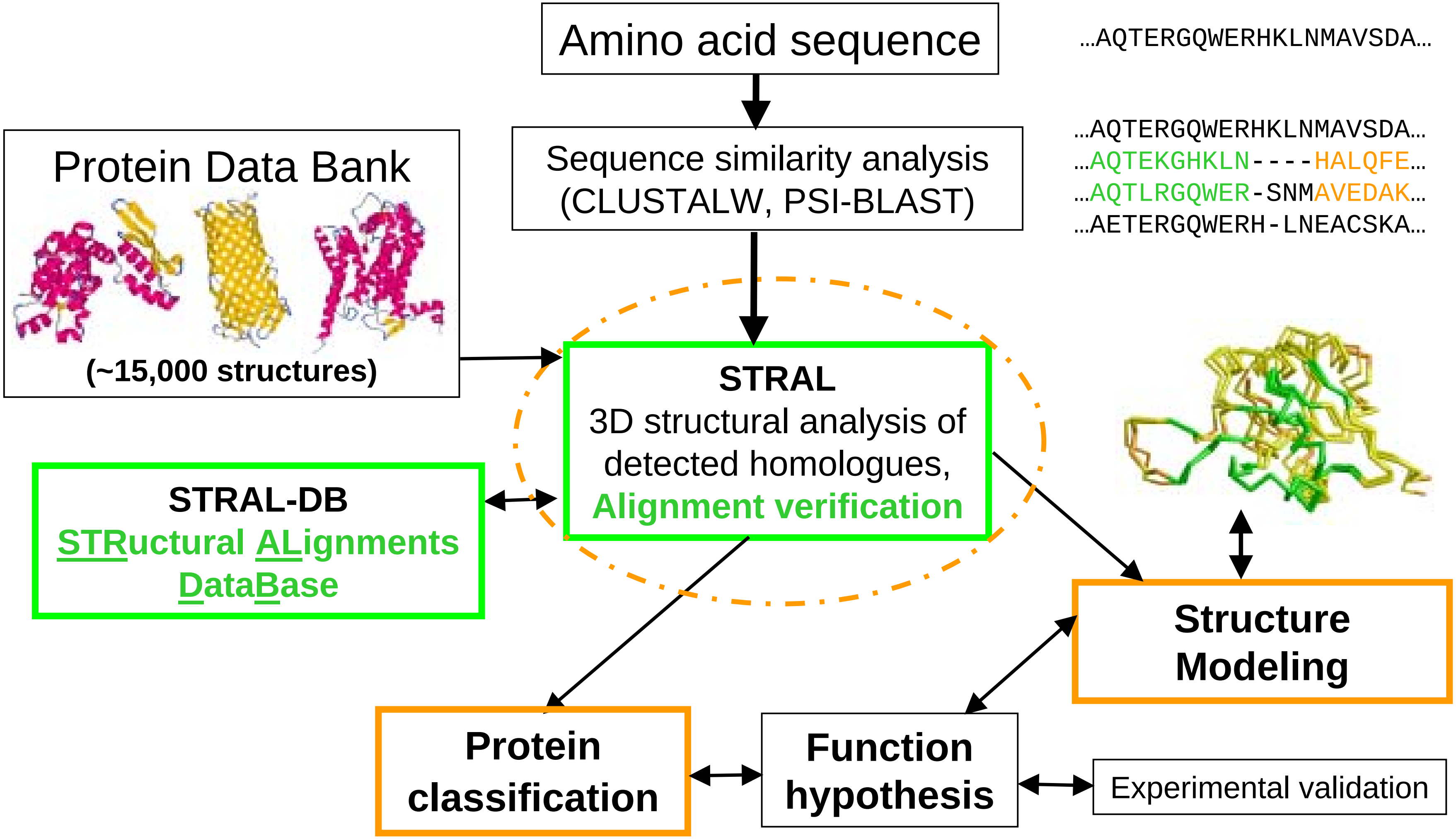
structure similarity varies

- Sequence based alignments may be wrong

and structure based validation is needed

Our target region

Our proposed work will provide computational improvements for protein classification



Amino acid sequence

…AQTERGQWERHKLNMAVSDA…

Sequence similarity analysis
(CLUSTALW, PSI-BLAST)

```
…AQTERGQWERHKLNMAVSDA…
…AQTEKGHKLN----HALQFE…
…AQTLRGQWER-SNMAVEDAK…
…AETERGQWERH-LNEACSKA…
```

Protein Data Bank

(~15,000 structures)

STRAL
3D structural analysis of detected homologues,
Alignment verification

STRAL-DB
STRuctural ALignments DataBase

Structure Modeling

Protein classification

Function hypothesis

Experimental validation

Structural analysis corrects sequence-based alignment

```
>1adl
CDAFVGTWKLVSSENFDDYMKEVGVGFA
TRKVAGMAKPNMIISVNGDLVTIRSEST
FKNTEISFKLGVEFDEITADDRKVKSII
TLDGGALVQVQKWDGKSTTIKRKRDGDK
LVVECVMKGVTSTRVYERA
```

1adl - 1cbi_A

N1	N2	DIST	N	Seq_Id	RMSD
131	136	5.0	127	37.80	2.01

```
>1cbi_A
PNFAGTWKMRSSENFDELLKALGVNAML
RKVAVAAASKPHVEIRQDGDQFYIKTST
TVRTTEINFKVGEGFEEETVDGRKCRSL
PTWENENKIHCTQTLLEGDGPKTYWTRE
LANDELILTFGADDVVCTRIYVRE
```

Structural alignment by STRAL

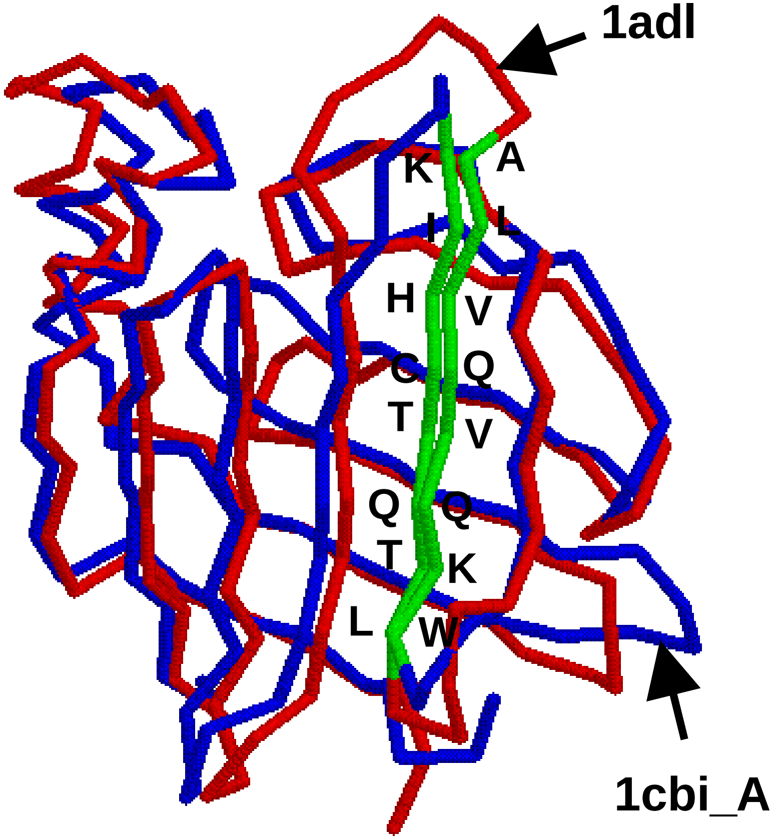
```
..........ALVQVQKW..........

..........KIHCTQTL..........
```

WRONG alignment by FASTA

```
.........ALVQ----VQKW.......
          \ \ \ \   / / / /
.........KIHCTQTL...........
```

WRONG alignment by PSI-BLAST

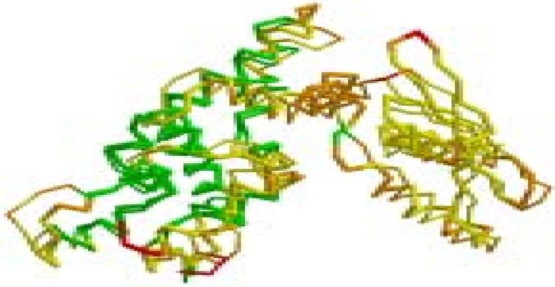
```
.........ALVQVQK----W.......
          \ \ \ \ \ \ \   /
.........KIHCTQTL...........
```



Our approach:
Multi-level method to determine similarities

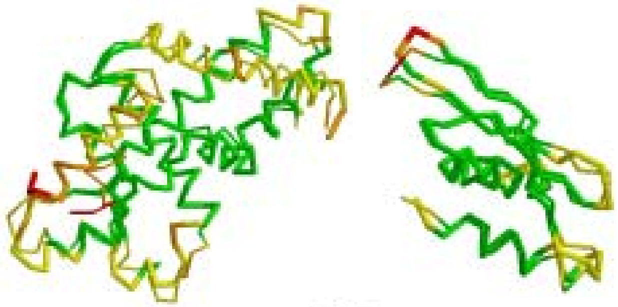
1. Discovery of overall structure similarity

   (typical state of the art)



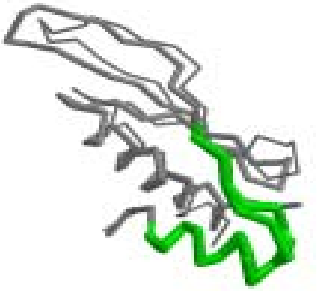
2. Analysis of similarities per domains

   (challenge starts here)



3. Refinement of the regions of local similarities within domains

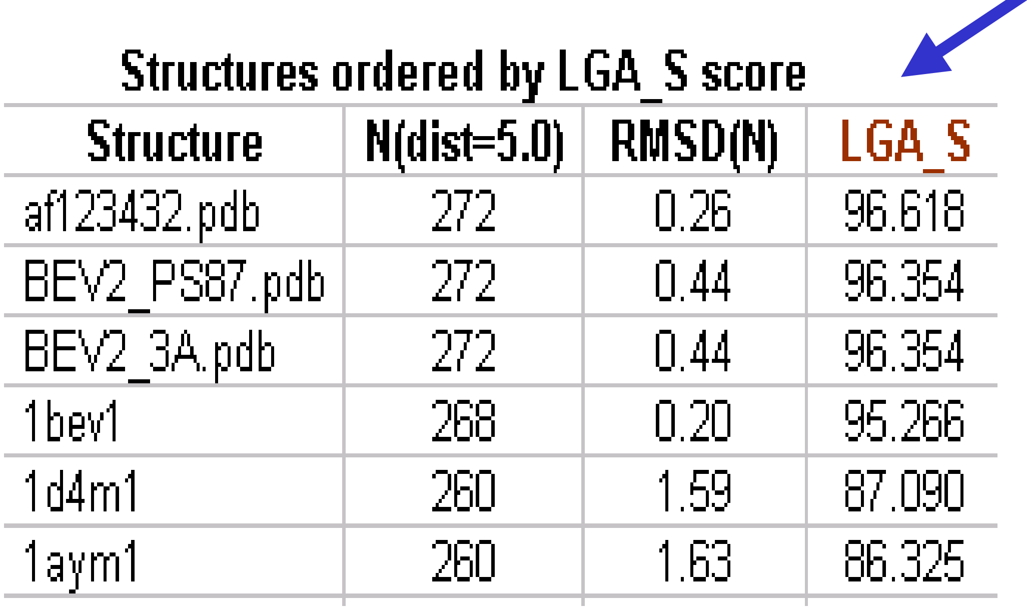
   - results assigned to each residue
   - retains high confidence anchoring determined at domain level



4. Evaluation function (scoring, ordering, alignment)

How can structures be compared?

Scoring function is key to identifying useful templates

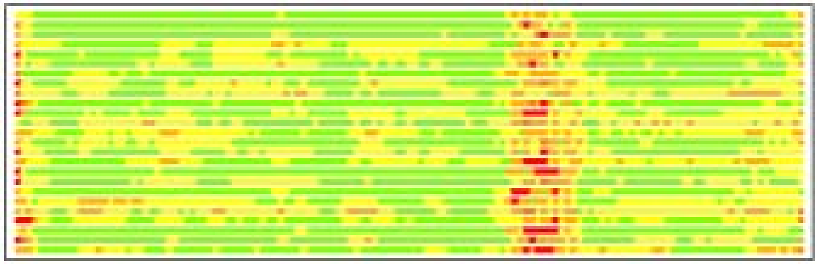


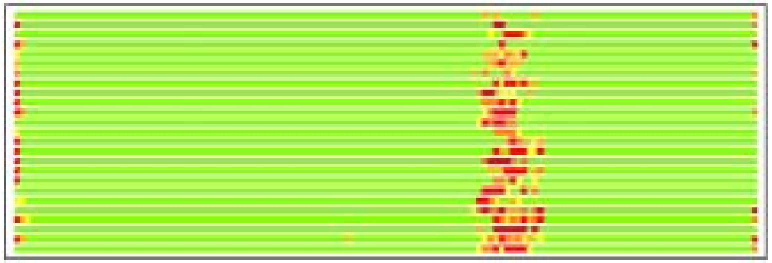
Structures ordered by LGA_S score

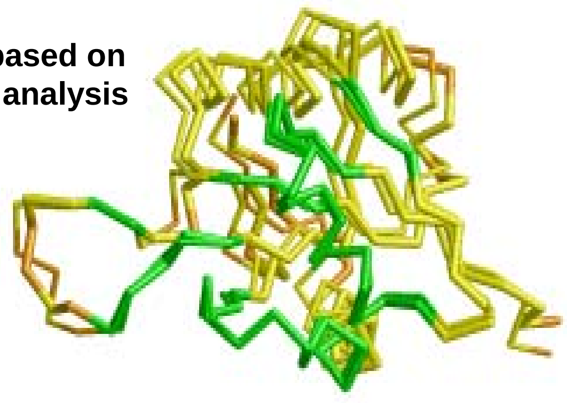
Structure	N(dist=5.0)	RMSD(N)	LGA_S
af123432.pdb	272	0.26	96.618
BEV2_PS87.pdb	272	0.44	96.354
BEV2_3A.pdb	272	0.44	96.354
1bev1	268	0.20	95.266
1d4m1	260	1.59	87.090
1aym1	260	1.63	86.325

Scoring function combines N (= number of amino-acids aligned) and RMSD (distance)

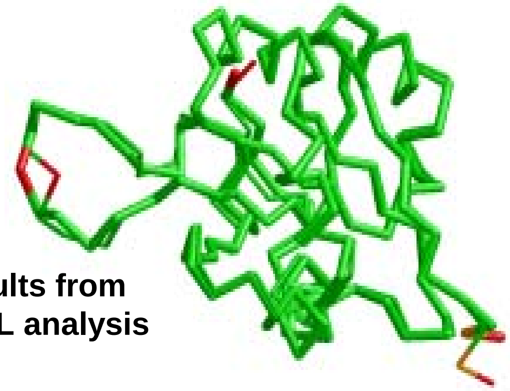
Early test of STRAL basic algorithm: handles “easy” case of high level of structure similarity







Results based on standard analysis



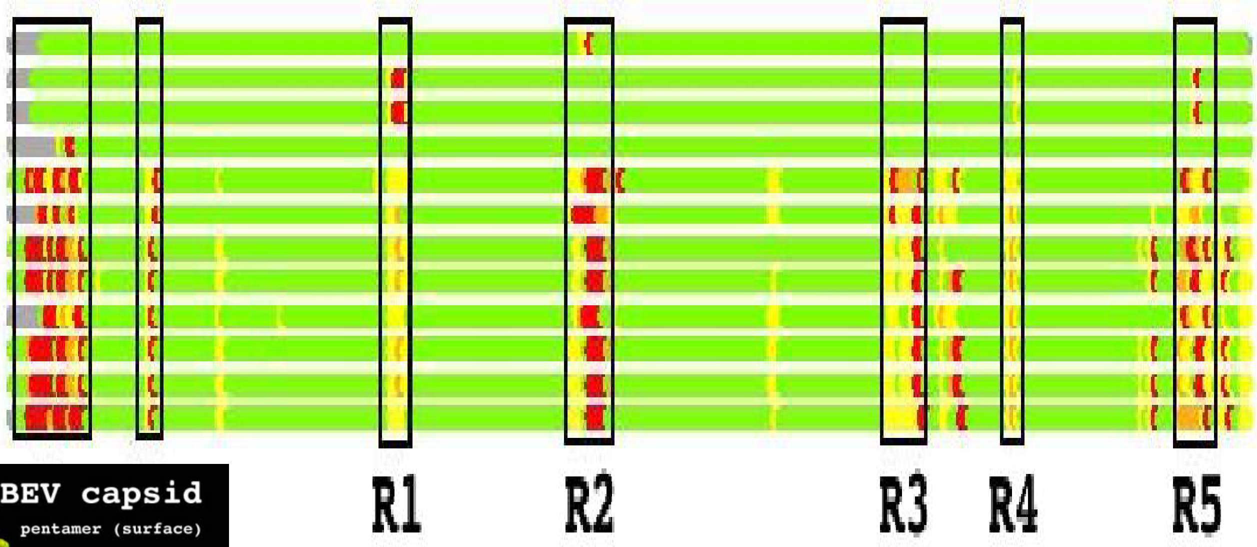
Results from STRAL analysis

In green – regions detected as very similar, in yellow – less similar, in red – not similar

Standard analysis does not distinguish the regions of similarity as clearly as our approach

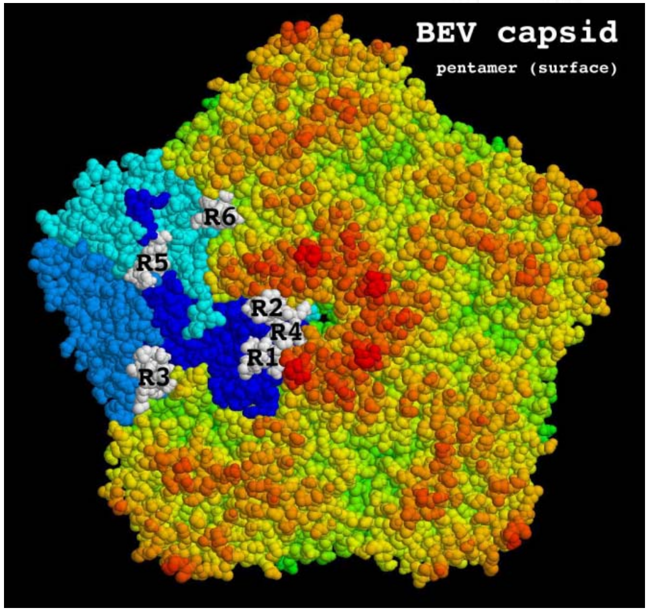
LGA structurally differentiates strains/species

Coat proteins from 12 enteroviruses



Structural similarity: green = high; yellow = moderate; red= little/none

Boxes: species- or strain-level differences in regions of biological interest



Regions of interest at or in “canyon” host receptor binding site

LGA can be used to identify structural epitopes as targets for detection, therapeutics, vaccines

The following individuals contributed to work summarized in this talk

Adam Zemla
Clinton Torres
Jason Smith
Carol Zhou
Tom Slezak
Beth Vitalis
Tom Kuczmarski
Marisa Lam
John Moult
Krzysztoff Fidelis
Tim Hubbard
Daniel Barsky
Dorota Sawicka